國立中正大學資訊工程系
常博愛個人作品集（部分）

目錄

1.畢業專題介紹（結尾附畢業專題設計海報，曾獲中正大學畢業展覽計算機理論組最佳人氣獎）
2.機器學習實作題目與報告（perceptron , Regression, GNN蘭花影像識別訓練，SVM，論文研讀）
3.演算法與資料結構專題程式設計（題目與算法，最短路徑(背包問題），MST，TSP，Merkle Tree）
4.數據科學概論專題題目與報告（降維比較，sampling，資料實作）
5.進階演算法程式專題與手寫作業（線性規劃—simplex;期末報告橢球法證明請見github）
6.線性代數課程規劃介紹與教學助理證明
7.網頁程式設計專題
8.社群媒體專題介紹（期中論文報告+期末台灣大選分析專題）
9.畢業專題海報
10.專題實驗第一學期學習總結（random walk，modularity，spectral clustering）
以上內容僅為大致介紹，同時皆有相應code與完整內容，歡迎到本人github查詢：
https://github.com/Brandon99650

畢業專題部分:

國立中正大學
National Chung Cheng University

基於AGM多標籤景觀聚類旅遊規劃

指導教授：李新林
學　　生：常博愛 蔡嘉祥

Introduction:

Object:在多標籤景點下自動為遊客推薦一條含有多種景點的短期旅遊路線。

Motivation:

1.旅游景點的非監督分群方法多基于單一標簽分群，特質單一

2. 逐個閱讀景點簡介爲游客帶來許多不便

3.為有觀光需求但時間不充裕的遊客自動規劃提高旅遊體驗

Data:

- 我們選擇從台灣交通觀光局獲得的台灣本土旅游資料集，包含約4500個景點
  - 資料特徵 ： 文字描述，所屬地區，經緯度

- 我們選擇開源API獲取兩地之間的行車距離（OpenStreetMap(OSMR) data）

Research Goals:

- RG1:如何實現多標籤分群?(part I)
  - Part I: 模糊分群（ fuzzy clustering）

- RG2:何種旅遊路線對有助於提高遊客體驗與實現方法?(part II)
  - Part II:路徑規劃
    1. 覆蓋種類多
    2. 標籤分佈平衡
    3. 路程時間短

Research Goals:

●RG1:如何實現多標籤分群?(part I)

- 非監督的，多標籤的
- 圖論算法
- AGM model
- 依據景觀描述的相似度分群

Part I:

BERT → KNN → AGM → GCN

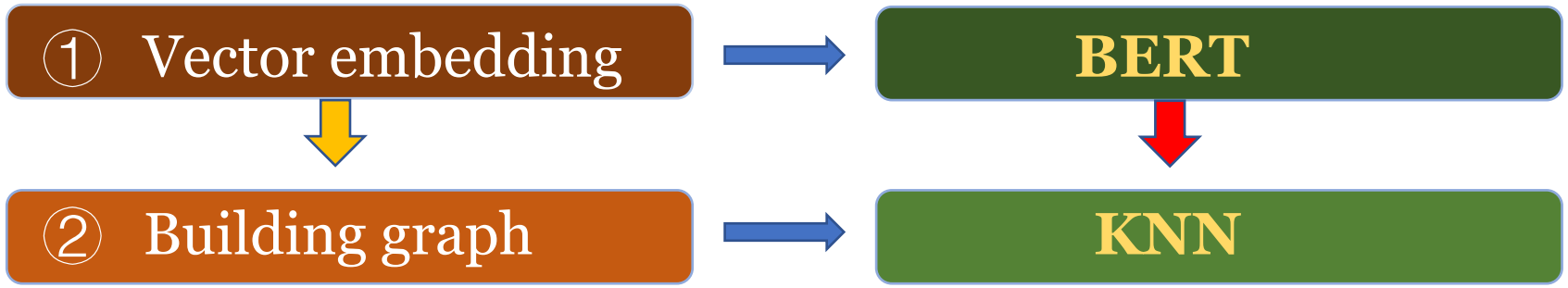
① Vector embedding → **BERT**

- **SentenceTransformer**
  - 預訓練任務： 文意相似度
    - Pretrained Models — Sentence-Transformers documentation (sbert.net)

- 目標： 使文意相似的簡介映射出來之向量cosine similarity 盡量高。

Part I:

BERT → KNN → AGM → GCN



根據映射出來之向量cosine similarity 建邊

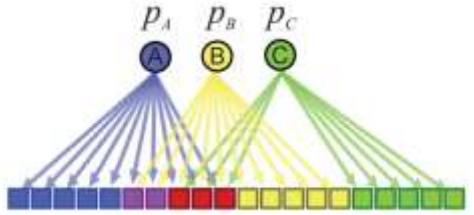
Part I:

BERT → KNN → AGM → GCN

① Vector embedding → BERT

② Building graph → KNN

③ Muti-label Clustering → AGM



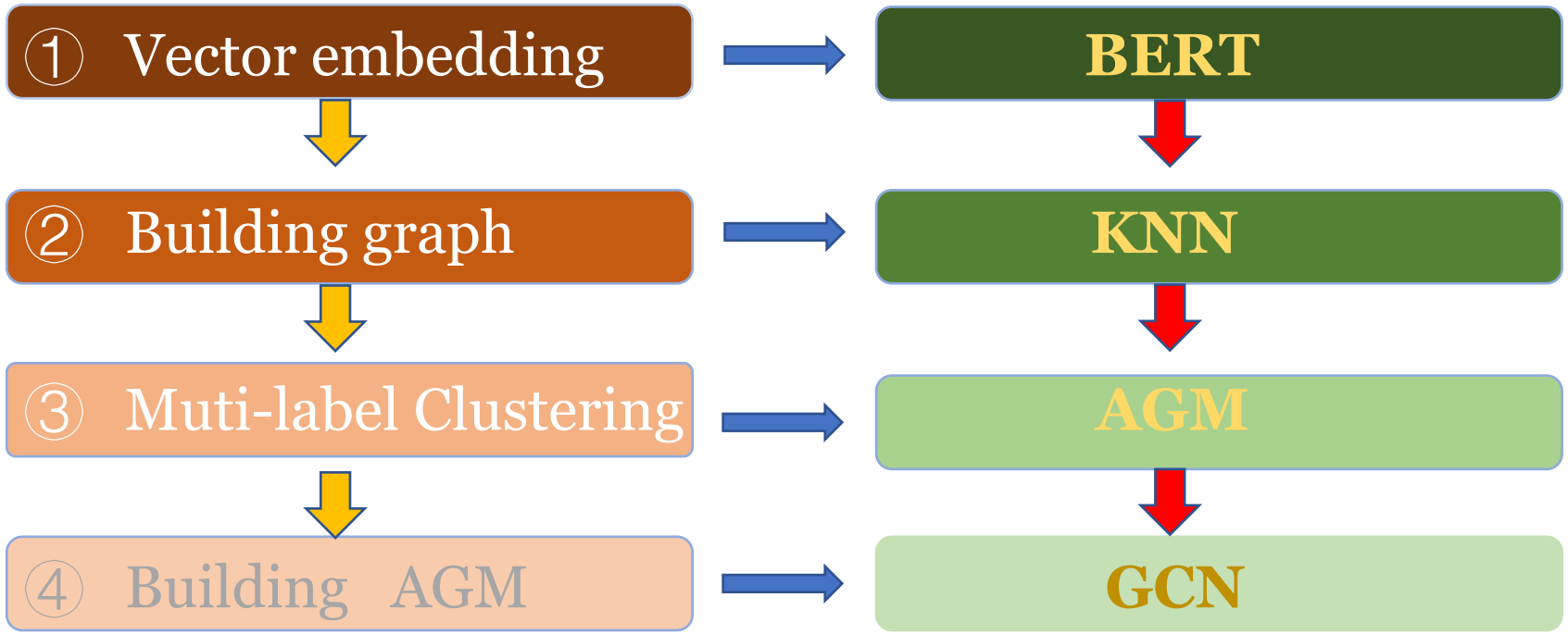
(a) Community Affiliation Network

(b) Overlap

- **AGM (Community-Affiliation Graph Model)** : $P(\mathrm{F}|\mathrm{G}) = -\sum_{(u,v)\in E} \log\left(1 - e^{-F_u F_v}\right) + \sum_{(u,v)\notin E} F_u F_v$
  - 目的:
    - 將每個節點嵌入到一個向量中，該向量表示它對每個社區的隸屬關係 (clustering)
    - $F^* = argmin_F(P(F|G))$
      $C_{nj} = \{c_i | c_i \geq thr., \forall c_i \in F^*(X_{C_{nj}}, A)\}$
    - $argmin(P(F|G))$:整張圖內相似屬性的點在同一社群的幾率高，不同屬性的點在同一社群的機率低

Part I:

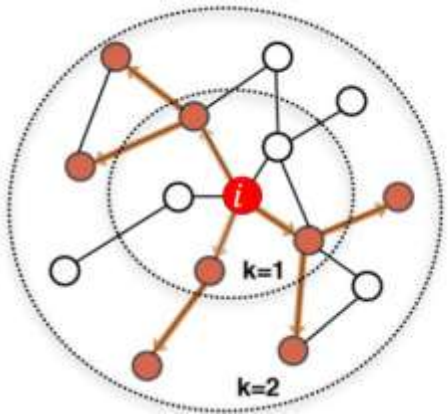
BERT → KNN → AGM → GCN

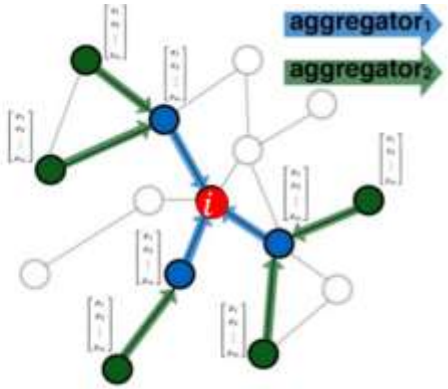


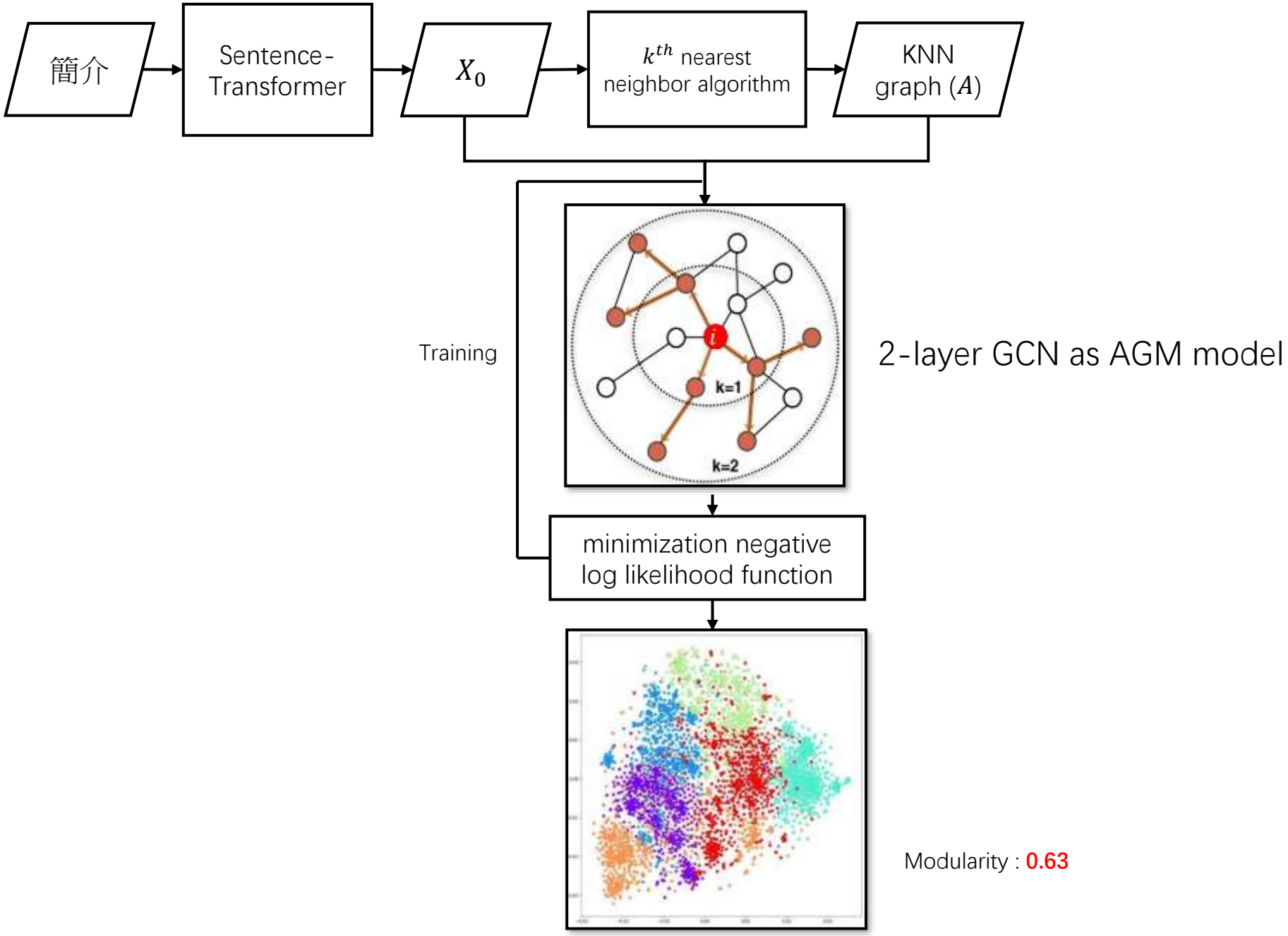
Part I:

BERT → KNN → AGM → GCN

- **GCN (Graph Convolution Neural Network) :**
  - 圖神經網路
  - 在每一層，每個點會得到鄰居的信息 (aggregate)
  - 多數使用: **2-layer** GCN model
    (over smooth problem)







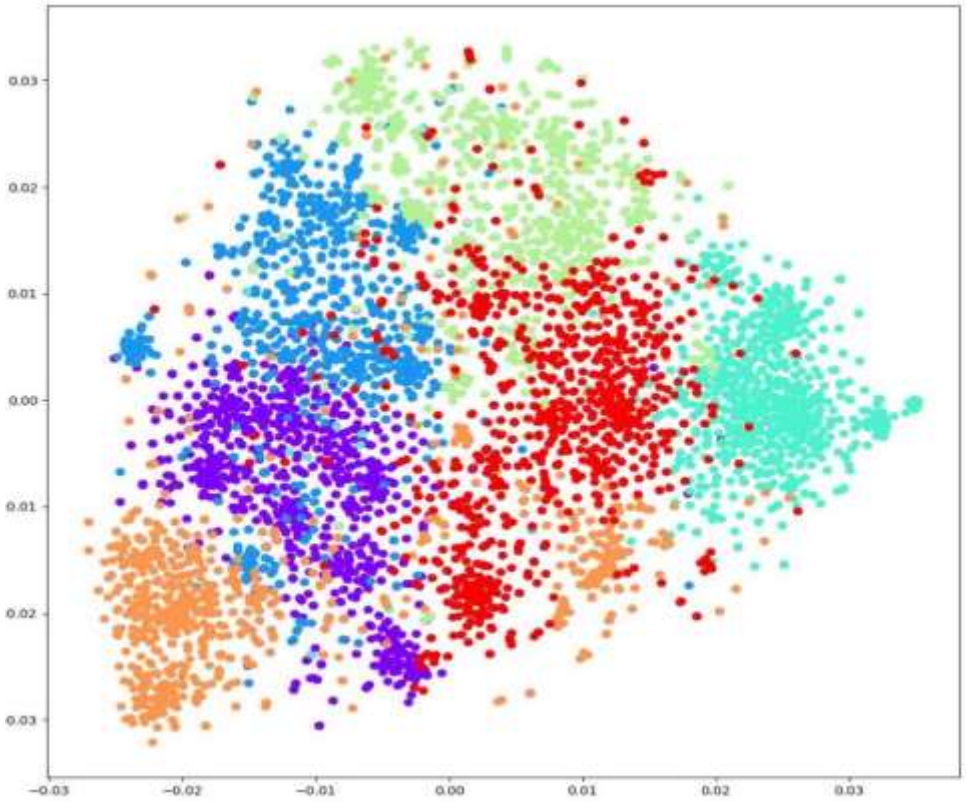
2-layer GCN as AGM model

Modularity : **0.63**

Part I:分群結果

BERT → KNN → AGM → GCN

分群(6群)結果(t-SNE降維)
modularity：0.63

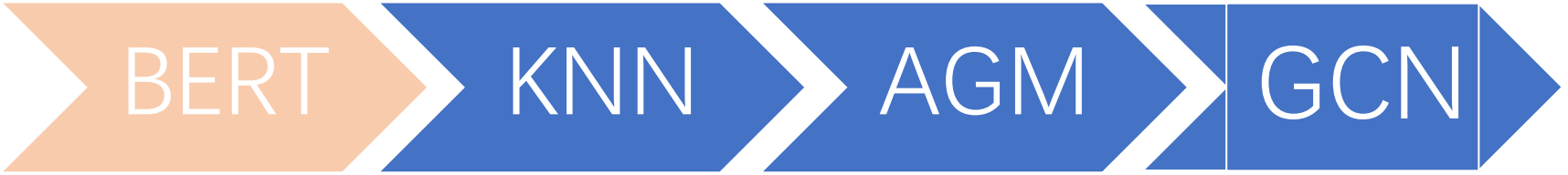


Modularity：一個判別community detection好壞的指標，越大則結果越好

$$Q = \frac{1}{2m}\sum_{i,j\in A}\left[\left(A_{ji} - \frac{d(i)d(j)}{2m}\right) * \delta(c_i, c_j)\right],$$

acceptable range: Q > 0.4

Part I:群內分析



使用TF-IDF關鍵字擷取技術，針對每群內之景點的簡介提取關鍵詞，並依據重複多次的關鍵詞對該群下有意義的標籤。

Tool : jieba (Python)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster5
自然探索	山野農情	文化信仰	藝術人文	海湖風光	歷史紀念

Research Goals:

- RG1:如何實現多標籤分群?(part I)
  - Part I: 模糊分群（ fuzzy clustering）

- RG2:何種旅遊路線對有助於提高遊客體驗與實現方法?(part II)
  - Part II:路徑規劃
    1. 覆蓋種類多
    2. 標籤分佈平衡
    3. 路程時間短

Part II:考量因素

- **標籤種類的覆蓋程度**（訪問過多少種類的標籤）
  - 越多種類越好
  - 最重要之因素

- **平衡度**:
  - 每種標籤被訪問的次數要盡量平均

$$score = n + e^{-\frac{Var(Labels)+\text{travel time} * w_t}{10}} \quad (w_t : 5)$$

- **行車時間**:
  - 短時間旅遊
  - 這段路程總行車時間越少越好

PartII:問題數學化

- **標籤種類的覆蓋程度:**
  - n := # **nonzero entry** of counting vector($Labels$) for each labels
  - n 越大越好
  - Dominate the good or bad.

- **平衡度:**
  - **Variance ($Labels$)**
  - Variance 越小，越平均，也就越好

0	1	2	3	4	5
#0	#1	#2	#3	#4	#5

Evaluation:

基于我們考量的因素，幫我們産生出的路徑評分

- $\boldsymbol{score = n + e^{-\frac{Var(Labels)+\text{travel time} * w_t}{10}}}$ $(w_t : 5)$
- 越大，此路綫越符合「多元」的定義

Part II：演算法

- 基本精神：
  - 將一定時間內開車可達到的地點作為候選人

    行車時間 ： routingpy (OSMR API)

  - 從候選人內優挑出能增加訪之標籤種類(n)的旅遊景點作為目的地可能之集和

  - 從其中挑出加入路徑後，variance 與行車時間和最小的景點作為下一站

Algorithm Generate diverse travel route

cur := source , $Path$:= {source}
$Labels$:= sourse.$Label$, total travel time := 0
while $|Path| < WantedNumber$ **do**
  $next := \emptyset$
  $Candidate \leftarrow$ traveltime$(cur, attraction_i) \leq$ time limitation,
    $\forall attraction_i \in$ dataset

  if # nonzero entry of $Labels$ can't be improved **then**
    $next \leftarrow \mathbf{argmin}_{c_i}(\mathbf{Var}(Label + c_i.Label)+\text{traveltime}(cur, c_i))$,
      $\forall c_i \in Candidate$
  else
    $Targets \leftarrow \mathbf{argmax}_{c_i}$ (# nonzero entry of $Labels + c_i.Label$),
      $\forall c_i \in Candidate$
    $next \leftarrow \mathbf{argmin}_{t_i}(\mathbf{Var}(Labels + t_i.Label)+\text{traveltime}(cur, t_i))$,
      $\forall t_i \in Targets$
  end if

  Update :
    $Path \leftarrow Path \cup next$, $Labels \leftarrow Labels + next.Label$
    total travel time $\leftarrow$ total travel time + traveltime$(cur, next)$
    $cur \leftarrow next$
end while
return $Path$, $Labels$, total travel time

Result:

路線1：

蘆竹溝觀光漁港➔東隆宮文化中心➔北門提溴塔➔北門自行車道

路線2：

台灣戲劇館➔吳沙故居➔柴圍佃圳水門公園➔礁溪溫泉地景廣場

蘆竹溝觀光漁港
[海湖風光]

東隆宮文化中心
[文化信仰、藝術人文]

北門提溴塔
[海湖風光、歷史紀念]

北門自行車道
[自然探索、海湖風光]

自然探索	山野農情	文化信仰	藝術人文	海湖風光	歷史紀念
1	0	1	1	3	1

經過種類 : 5　　Variance : 0.805
行車總時長: 0.18 hrs (10mins)
Score: 5.176 (70%)

台灣戲劇館
[藝術人文]

吳沙故居
[文化信仰、歷史紀念]

柴圍佃圳水門公園
[自然探索、山野農情]

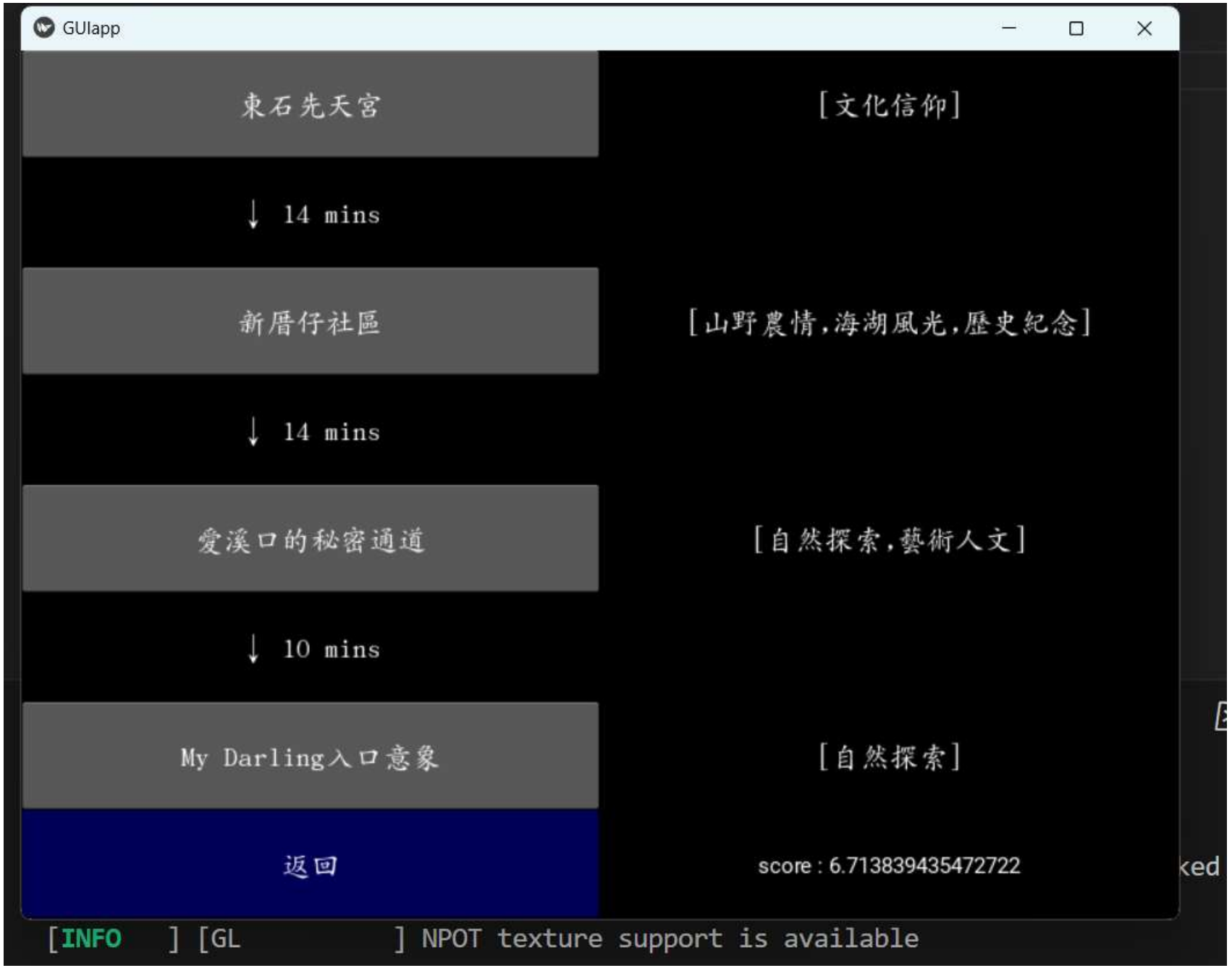
礁溪溫泉地景廣場
[山野農情、海湖風光]

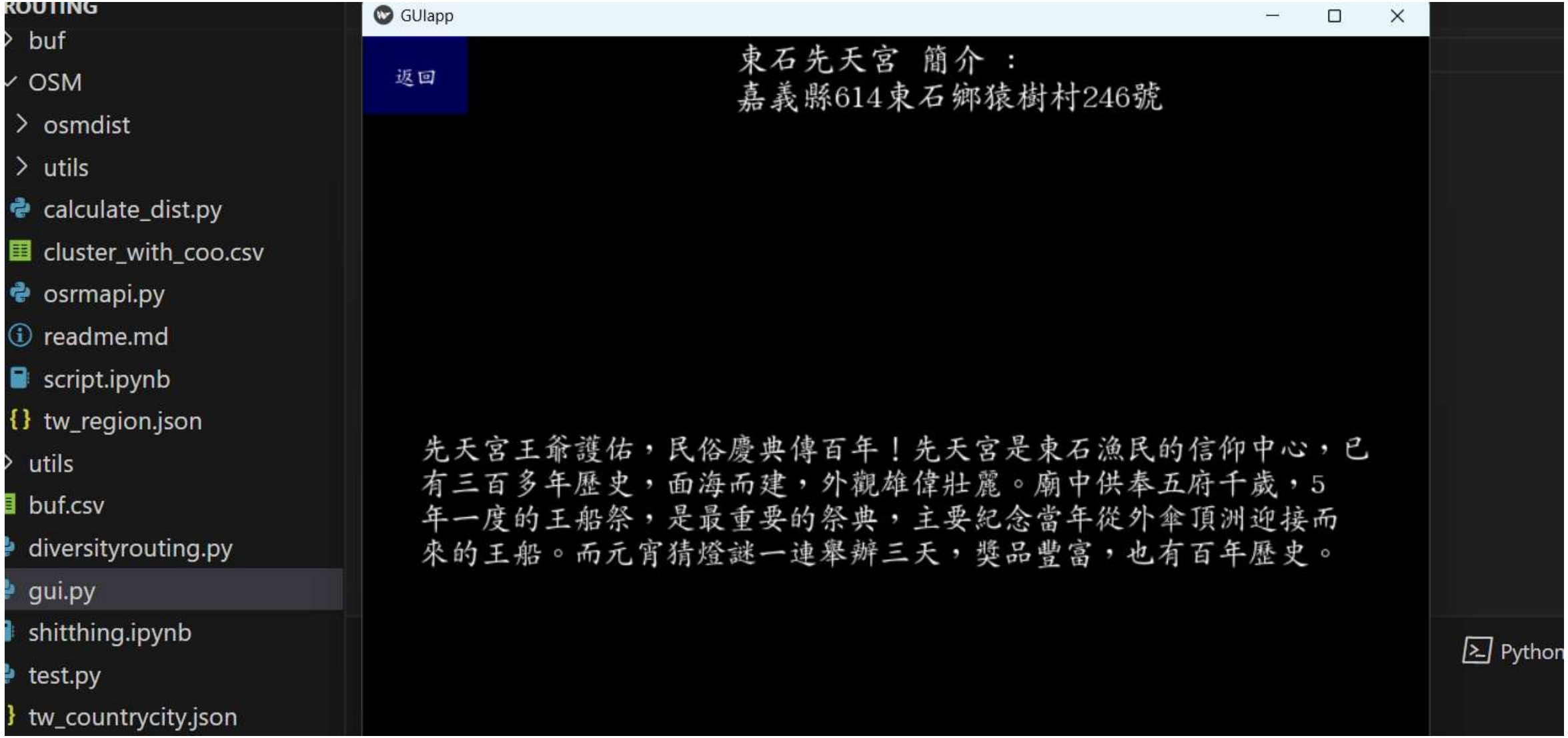
自然探索	山野農情	文化信仰	藝術人文	海湖風光	歷史紀念
1	2	1	1	1	1

經過種類 : 6　Variance : 0.138
行車總時長: 0.35 hrs (21mins)
Score: 6.154　(86%)

可視化GUI:


ROUTING
buf
OSM
osmdist
utils
calculate_dist.py
cluster_with_coo.csv
osrmapi.py
readme.md
script.ipynb
tw_region.json
utils
buf.csv
diversityrouting.py
gui.py
shitthing.ipynb
test.py
tw_countrycity.json
ukai.ttc
GUIapp
標籤
自然探索
山野農情
文化信仰
藝術人文
海湖風光
歷史紀念
兩地時限(min):
default:30
城市
直接指定起點:
番路鄉
梅山鄉
竹崎鄉
阿里山鄉
中埔鄉
大埔鄉
水上鄉
鹿草鄉
太保市
朴子市
東石鄉
六腳鄉
新港鄉
民雄鄉
大林鎮
溪口鄉
義竹鄉
布袋鎮
返回
GO!
Python


GUIapp
東石先天宮
[文化信仰]
↓ 14 mins
新厝仔社區
[山野農情,海湖風光,歷史紀念]
↓ 14 mins
愛溪口的秘密通道
[自然探索,藝術人文]
↓ 10 mins
My Darling入口意象
[自然探索]
返回
score : 6.713839435472722
[INFO] [GL] NPOT texture support is available


ROUTING
buf
OSM
osmdist
utils
calculate_dist.py
cluster_with_coo.csv
osrmapi.py
readme.md
script.ipynb
tw_region.json
utils
buf.csv
diversityrouting.py
gui.py
shitthing.ipynb
test.py
tw_countrycity.json
GUIapp
返回
東石先天宮 簡介 ：
嘉義縣614東石鄉猿樹村246號
先天宮王爺護佑，民俗慶典傳百年！先天宮是東石漁民的信仰中心，已有三百多年歷史，面海而建，外觀雄偉壯麗。廟中供奉五府千歲，5年一度的王船祭，是最重要的祭典，主要紀念當年從外傘頂洲迎接而來的王船。而元宵猜燈謎一連舉辦三天，獎品豐富，也有百年歷史。
Python

Conclusion:

- 透過我們設計的這項推薦演算法，我們給使用者提供了豐富而多元的短路程旅游參考路綫。
- 未來展望：
  ① 如果可以針對資料集訓練出專屬的NLP模型，或許可以更進一步改善旅游景點標籤的合理性，進而改善整體結果。
  ② 未來可考慮優化模型在分群過程中得到每個景點屬於各個標籤的機率，並以此作為依據為路徑打分分析得到該路徑的優勢及強度。

●RG1:How to multi-label clustering ?(part I)

BERT > KNN > GCN > AGM

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster5
自然探索	山野農情	文化信仰	藝術人文	海湖風光	歷史紀念

●RG2:What kind of travel plan will be useful for users and how to complete it ?(part II)

在優先考慮經過景點種類越多越好的情況下同時考量標籤分佈盡量平均和行駛路程盡量短

Reference

- **[1]Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach**
  - Yang, J., & Leskovec, J. (2013, February). Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 587-596).

- **[2]Overlapping Community Detection with Graph Neural Networks**
  - Shchur, O., & Günnemann, S. (2019). Overlapping community detection with graph
  - neural networks. *arXiv preprint arXiv:1909.12201*.

- Source Code: https://github.com/Brandon99650/Graduation-Project